Audio Based Object Detection System: A Comprehensive Survey

Aishwarya Pralhad Chakrapani
Department of Computer Engineering
PVPIT
aishwaryachakrapani12@gmail.com

Aryan Mayur Dhuru
Department of Computer Engineering
PVPIT
aryandhuru49@gmail.com

Shraawani Prakash Lattoo
Department of Computer Engineering
PVPIT
shraawani.lattoo@gmail.com

Parth Samir Dalvi
Department of Computer Engineering
PVPIT
parthdalvi3.1415@gmail.com

Rohit Sanjay Shahane
Department of Computer Engineering
PVPIT
rohit.shahane2002@gmail.com

---***---

Abstract - This survey addresses visual impairment by leveraging advanced technology to empower independent navigation for visually impaired individuals. It emphasizes the importance of tailored systems and highlights successful studies implementing innovative solutions. Incorporating diverse technologies like object detection, NLP, and information retrieval, the project explores deep learning algorithms such as CNN, RNN, and YOLO, alongside audio-based input/output integration. It recommends utilizing APIs like Google Text-to-Speech and Python libraries for efficient implementation, aiming to enhance system functionality and accessibility. Ultimately, the survey aims to aid in selecting appropriate tools and methodologies for developing audio-based object detection systems, contributing to ongoing efforts in supporting visually impaired individuals.

***Key Words*:** Object Detection, Natural Language Processing, Audio Data, YOLO, Real-Time identification, Visually Impaired, Image Captioning.

INTRODUCTION

Visual impairment presents significant obstacles for individuals navigating their surroundings and accessing information independently. Globally, the World Health Organization (WHO) estimates that around 253 million people are affected by vision impairment, with 36 million categorized as blind. While traditional aids like canes and guide dogs offer some support, advancements in technology offer a chance to transform how visually impaired individuals interact with their environment. This project seeks to overcome the challenges faced by visually impaired individuals by harnessing advanced technologies, particularly in artificial intelligence (AI) and deep learning. Through techniques such as object detection, natural language processing (NLP), and audio-based interfaces, our goal is to develop a system that empowers visually impaired individuals to navigate with confidence and independence. The envisioned system will comprise a blend of hardware and software components, incorporating sensors, cameras, and a central processing unit equipped with deep learning algorithms. These algorithms will analyze real-time environmental data, identifying objects and obstacles and providing auditory feedback to guide user decisions during navigation.

Beyond enhancing mobility, the system will prioritize improving access to information for visually impaired individuals. By integrating with existing platforms such as GPS and online databases, users will gain access to real-time information about their surroundings. Ultimately, this project aims not only to offer practical assistance to visually impaired individuals but also to foster inclusivity and accessibility within our society. Leveraging the latest advancements in AI and deep learning, we aim to empower visually impaired individuals to lead more independent and fulfilling lives.

RECENT WORK

1. “Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3”, Mansi Mahendru, Sunjay Kumar Dubey.

Object recognition is an essential component of computer vision. It has applications in autonomous vehicles, robotics, and assisting people with visual impairments. In this paper, we present a proposal for a system that utilizes Yolo and Yolo_v3 algorithm to identify multiple daily objects and provide voice commands to users. The system utilizes Google text to speech(gTTS) and Pygame python module for audio feedback.We tested this system on a dataset with over 200,000 images to determine how well both algorithms interact with a webcam in various situations.The aim of the system is to assist visually impaired people to navigate unfamiliar environments efficiently using deep learning algorithms and audio feedback. The aim of this research is to improve accessibility and usability for people with visual impairment through advances in computer vision and machine learning technologies.

2. “Real-Time Object Detection for Visually Challenged People”, Sunit Vaidya; Naisha Shah; Niti Shah; Radha Shankarmani

Vision is a vital sense for living beings, but millions worldwide grapple with visual impairment, posing challenges in navigation, information access, and communication. The aim of the proposed project is to transform the visual landscape into an auditory one, alerting blind individuals to objects in their vicinity. Through a real-time object detection system, visually impaired individuals can navigate independently, with image processing and machine learning identifying objects via a camera and conveying their location through audio descriptions. Existing methods suffer from accuracy and performance issues, highlighting the need for enhanced approaches. Therefore, the project focuses on improving accuracy, performance, and usability to offer visually impaired individuals a reliable means of autonomous navigation, thereby enhancing their quality of life.

3. “Object Captioning and Retrieval with Natural Language”, Anh Nguyen, Quang D. Tran, Thanh-Toan Do; Ian Reid, Darwin G. Caldwell, Nikos G. Tsagarakis

We tackle the challenge of integrating vision and language to comprehend objects in greater detail. Our approach hinges on utilizing object descriptions to enhance understanding. We introduce two novel architectures aimed at addressing two interrelated tasks: object captioning and natural language-based object retrieval. Object captioning involves detecting an object and generating its description simultaneously, while object retrieval entails localizing an object based on a given query. We demonstrate the effectiveness of our methods, employing hybrid CNN-LSTM networks, on a newly developed dataset. Our experimental results show significant improvements over recent approaches, delivering both detailed object comprehension and efficient inference. The source code for our methods will be released for further exploration.

4. “Real-Time Object Detection and Audio Feedback for the Visually Impaired”, Ayan Ravindra Jambhulkar; Akshay Rameshbhai Gajera; Chirag Manoj Bhavsar; Shilpa Vatkar

The lives of visually impaired individuals are often fraught with challenges, particularly in independently identifying and navigating their surroundings. The utilization of computer vision-based object detection techniques has emerged as a promising avenue for assisting the visually impaired by swiftly detecting and categorizing objects in real-time. This study introduces a pioneering real-time object detection and audio feedback system tailored to empower visually impaired individuals in recognizing and navigating their environment autonomously. Our method harnesses the YOLO_v3 algorithm in tandem with the MS COCO dataset to rapidly detect and classify objects, subsequently providing corresponding audio feedback. The system implementation seamlessly integrates the gTTS (Google Text-to-Speech) API to generate audio feedback, leveraging advanced audio processing techniques and deep learning algorithms. Evaluation conducted on a comprehensive dataset yielded a remarkable average detection accuracy of 90%. This groundbreaking system presents a pragmatic and impactful solution for enhancing accessibility and independence for visually impaired individuals, underscoring the effectiveness of employing sophisticated deep learning algorithms and datasets in real-time object detection and audio feedback systems.

5. “Visual object detection using audio data”, Rajnish Kumar Chaturvedi, Dinesh Prasad Sahu, Manoj Kumar Tyagi, Manoj Diwakar, Prabhishek Singh, Achyut Shankar and V E Sathishkumar

In today's rapidly evolving landscape of Internet of Things (IoT) and Machine Learning (ML), object detection stands out as a significant application. This technology, which identifies semantic entities in digital images and videos such as humans, vehicles, and buildings, benefits greatly from visual data acquired through cameras. However, it often grapples with limitations stemming from a confined Field of View. To address this issue, this study integrates audio data into the object localization process. By employing a microphone to estimate the angular position of sound-emitting objects, this approach supplements camera-based detection. Initially, objects within the camera's range are identified and tracked using optical flow. However, when these objects move out of view, the microphone aids in determining their angular position. This ensures continuous localization of objects, facilitating subsequent ML-driven tasks like facial recognition using IoT devices.

6. Lightweight Real-time Object Detection System Based on Embedded AI Development Kit -Junjie Li; Xinsen Zhou; Qianqiu Wang; Xianlu Luo

The focus of lightweight object detection is on optimizing model performance for real-time applications by minimizing model size without sacrificing detection accuracy. This system enhances the MobileNet-SSD object detection algorithm and utilizes standard datasets for training and evaluation. Through channel pruning, model parameters are significantly reduced while maintaining accuracy, resulting in a compression ratio of 1.17:1. This reduces the model's memory footprint and yields a 44% improvement in detection speed. The trained model is then deployed on the EAIDK-610 embedded artificial intelligence development kit to process and detect real-time video content. This approach enables practical detection tasks in scenarios with limited computing resources. Overall, the system's advancements in reducing model size and enhancing detection speed facilitate its deployment across various specific applications requiring efficient object detection within constrained computational environments.

7. Attentive Layer Separation for Object Classification and Object Localization In Object Detection - Jung Uk Kim and Yong Man Ro

In the domain of computer vision, object detection stands as a pivotal area. This encompasses tasks like classifying and localizing objects. Traditional deep learning models for object detection typically rely on shared networks to produce feature maps. However, while object classification zeroes in on the most distinguishing features of an object within the map, object localization demands a broader focus spanning the entire object area. This paper introduces a fresh approach to object detection that acknowledges this distinction. It comprises two primary components: an Attention network for generating task-specific attention maps and a Layer separation module for distinct task estimation layers. Rigorous experiments conducted on the PASCAL VOC and MS COCO datasets showcased the superior performance of the proposed object detection network over state-of-the-art methods.

8. “Object Detection with Audio Comments using YOLO v3”, Sneha Gupta, Suchismita Chakraborti, R Yogitha, G Mathivanan

Vision impairment profoundly impacts individuals' lives, hindering their independence, mobility, and safety as they navigate obstacles and potential tripping hazards. Recent studies by the World Health Organization reveal that approximately 2.2 billion people worldwide experience visual impairment, with a significant portion aged over 50 years. This paper focuses on the transformative potential of a simple Object Detection model, utilizing Python libraries and APIs, to aid visually impaired individuals in overcoming these challenges. The proposed methodology introduces a Machine Learning model designed to detect object locations surrounding the individual and provide voice feedback using the Google Text-to-Speech API. Training the model utilizes the MS-COCO dataset to improve its effectiveness in recognizing diverse objects with various aspects.

9. Guiding Visually Impaired People to Find an Object by Using Image to Speech over the Smart Phone Cameras. -Tayyip Mert Denizgez; Orçun Kamiloğlu; Seda Kul; Ahmet Sayar.

Visual impairment is a pressing global health issue, affecting one in four individuals, a figure exacerbated by increased technology use. Despite numerous existing solutions, many are either costly or ineffective. This paper introduces a groundbreaking system designed to assist visually impaired individuals in object localization. Utilizing image-to-speech technology through smartphone cameras, the system offers voice-guided directions. A pivotal innovation is the use of the user's hand as a reference point. By detecting the user's hand and identifying target objects within the camera's view, the system guides users to the object's location using image-to-speech techniques. This method calculates the target object's position relative to the user's hand using deep learning and image processing, with outcomes communicated via speech. Object detection employs a Convolutional Neural Network (CNN) model based on the Single Shot MultiBox Detector (SSD) approach, chosen for its superior accuracy and frame rate compared to other models on smartphones. The TensorFlow-Lite model, trained on the Common Objects in Context (COCO) dataset, encompasses ninety-one object classes.

10. Voice Assisted Object Detection for Visually Impaired, Kg Hitaish; Vani Krishnaswamy; M Vishnu; B Mahima, 2023

Visual impairment can significantly impact an individual's autonomy, employment opportunities, and daily activities. Globally, there are more than 2.2 billion people affected by various degrees of visual impairment, making tasks such as reading, writing, and navigation challenging. This can result in a lower quality of life and feelings of isolation and melancholy. Several deep learning models for object detection, grounded in computer vision, are available. Our goal is to incorporate the MobileNet SSD architecture into a navigation system that offers swift and precise object detection to aid individuals with visual impairments. Our project emphasizes mobility, featuring a lightweight architecture optimized for mobile and embedded devices with limited CPU capabilities. Furthermore, we have devised algorithms to deliver directional guidance and distance information, complemented by auditory feedback. In forthcoming iterations, this system can be adapted and implemented across diverse domains.

11. Deep learning-based object detection and surrounding environment description for visually impaired people Raihan Bin Islam, Samiha Akhter, Faria Iqbal, Md. Saif Ur Rahman, Riasat Khan, 2023

This paper introduces an affordable assistive system designed for visually impaired individuals, employing advanced deep learning methods for obstacle detection and environmental interpretation. By leveraging TensorFlow's object detection API and SSDLite MobileNetV2, the system can identify objects in real-time video streams captured by a Raspberry Pi camera. Audio feedback is generated using Google's text-to-speech technology, PyAudio, and speech recognition. Unlike traditional white canes, this device, integrated into a head cap, offers superior efficiency. Additionally, a special "ambiance mode" utilises transfer learning on weather data to provide detailed descriptions of surroundings. Performance assessments conducted on both desktop and Raspberry Pi systems demonstrate the system's accuracy using various metrics such as detection accuracy, mean average precision, frame rate, confusion matrix, and ROC curve. This cost-effective solution aims to significantly improve the daily experiences of visually impaired individuals.

12. Image Captioning using Convolutional Neural Networks and Recurrent Neural Network, Rachel Calvin, Shravya Suresh

Image captioning involves crafting precise textual explanations for online images, harnessing the power of Computer Vision and Natural Language Processing. This process employs advanced deep learning methods, utilizing Convolutional Neural Networks (CNNs) to extract features and Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) networks, to generate captions. The Flickr8_k Dataset is utilized for training these models. CNNs extract visual features from images, while RNNs generate coherent captions based on these extracted features. By combining CNNs and RNNs, the model effectively bridges the gap between visual content and natural language, enabling the production of accurate image descriptions.

13. YOLOv7: Trainable bag-of-freebies sets new state-of the-art for real-time object detectors, Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao

The steady advancements in real-time object detection have given rise to significant research fields in architecture and training optimization. To address these issues, a trainable bag-of-freebies approach is proposed, which combines flexible training tools with optimized architectures and compound scaling methods. YOLOv7 is a significant advancement as it surpasses existing detectors in accuracy and speed, offering a broad FPS range of 5 to 120 FPS. It stands out as the most accurate real-time detector on the GPU V100, with an Average Precision (AP) of 56.8%, among those operating at 30 frames per second or higher. This solution shows how cutting-edge techniques can be combined to achieve real-time object detection performance breakthroughs.

MODEL ARCHITECTURE

The project aims to develop a sophisticated system for object detection and audio integration, catering to both visually impaired and sighted users. It involves the utilization of state-of-the-art object detection algorithms such as YOLO or Faster R-CNN, seamlessly integrated with deep learning frameworks like TensorFlow or PyTorch. These algorithms are complemented by precise bounding box regression mechanisms to ensure accurate object localization. The audio integration module handles user inputs and delivers natural, intuitive responses through text-to-speech (TTS) systems, ensuring compatibility with standard audio input/output devices. The user interface is thoughtfully designed with accessibility features, offering intuitive audio-driven interactions and leveraging natural language processing for user commands. To enhance robustness, the system implements rigorous error handling mechanisms and secure communication protocols. Scalability is addressed through the utilization of distributed computing solutions, while compatibility is maintained across various platforms and devices. A comprehensive testing framework is established to evaluate system efficacy, alongside thorough documentation and deployment procedures to facilitate seamless implementation. Modularity enables future updates and enhancements, while monitoring tools provide insights into system performance. The system ensures regulatory compliance with data protection and accessibility standards, ensuring user privacy and inclusivity.

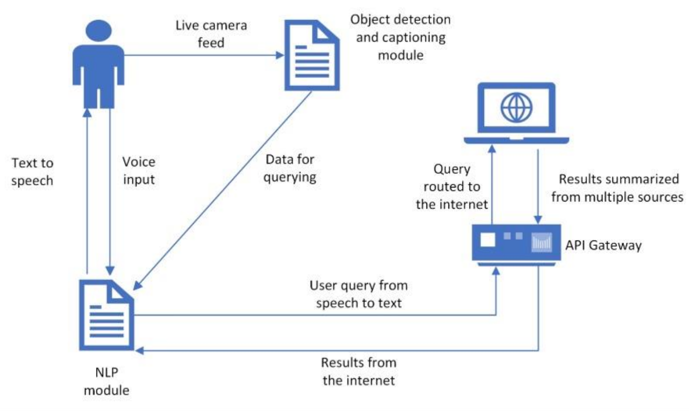
Live camera feed
Object detection and captioning module
Text to speech
Voice input
Data for querying
Query routed to the internet
Results summarized from multiple sources
API Gateway
User query from speech to text
Results from the internet
NLP module

Fig-1: Model

CONCLUSION

In conclusion, this project represents a significant step forward in leveraging technological advancements to address the challenges faced by visually impaired individuals. By harnessing the power of object detection, natural language processing, and deep learning algorithms such as CNN, RNN, and YOLO, along with the integration of audio-based input/output systems, we have created a comprehensive framework aimed at empowering visually impaired individuals to navigate their surroundings independently.

Through the review and justification of select studies, we have demonstrated the effectiveness of various methodologies in addressing the multifaceted nature of the problem. Additionally, by proposing the utilization of APIs such as Google Text-to-Speech and Python libraries, we have provided practical solutions to facilitate the implementation and enhance the functionality of the system.

Overall, this project serves as a valuable guide for researchers, developers, and stakeholders involved in the development of audio-based object detection systems for the visually impaired. By contributing to ongoing efforts in this field, we hope to make meaningful strides towards creating a more inclusive and accessible world for all individuals, regardless of visual impairment.

REFERENCES

1. Mahendru, M. and Dubey, S.K. (2021) Real Time Object Detection with Audio Feedback using Yolo vs. Yolo_v3, IEEE Xplore. https://doi.org/10.1109/Confluence51648.2021.9377064.

2. Vaidya, S. et al. (2020) Real-Time Object Detection for Visually Challenged People, IEEE Xplore. https://doi.org/10.1109/ICICCS48265.2020.9121085.

3. Nguyen, A., Tran, Q. D., Do, T., Reid, I., Caldwell, D. G., & Tsagarakis, N. G. (2019). Object Captioning and Retrieval with Natural Language. https://doi.org/10.1109/iccvw.2019.00316.

4. Jambhulkar, A. R., Gajera, A. R., Bhavsar, C. V., & Vatkar, S. (2023). Real-Time Object Detection and Audio Feedback for the Visually Impaired. Ieee. https://doi.org/10.1109/asiancon58793.2023.10269899.

5. Chaturvedi, R. K., Sahu, D. P., Tyagi, M. K., Diwakar, M., Singh, P., Shankar, A., & Sathishkumar, V. E. (2023). Visual object detection using audio data. Journal of Physics: Conference Series, 2664(1), 012006. https://doi.org/10.1088/1742-6596/2664/1/012006.

6. Li, J., Zhou, X., Wang, Q., & Luo, X. (2022). Lightweight real-time object detection system based on embedded AI development kit. 2022 International Conference on Machine Learning and Intelligent Systems

Engineering (MLISE). https://doi.org/10.1109/mlise57402.2022.00010.

7. Kim, J.U. and Man Ro, Y. (2019) 'Attentive layer separation for object classification and object localization in object detection', 2019 IEEE International Conference on Image Processing (ICIP) doi:10.1109/icip.2019.8803439.

8. Gupta, S., Chakraborti, S., Yogitha, R., & Mathivanan, G. (2022). Object Detection with Audio Comments using YOLO v3. 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC). https://doi.org/10.1109/icaaic53929.2022.9792755.

9. Denizgez, T.M. et al. (2021) 'Guiding visually impaired people to find an object by using image to speech over the smart phone cameras', 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). doi:10.1109/inista52262.2021.9548122.

10. Hitaish, K. et al. (2023) 'Voice assisted object detection for visually impaired', 2023 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). doi:10.1109/conecct57959.2023.10234781.

11. Islam, R.B. et al. (2023) 'Deep learning-based object detection and surrounding environment description for visually impaired people', Heliyon, 9(6). doi: 10.1016/j.heliyon. 2023.e16924.

12. Calvin, R., & Suresh, S. (2021). Image Captioning using Convolutional Neural Networks and Recurrent Neural Network. Ieee. https://doi.org/10.1109/i2ct51068.2021.9418001.

13. Wang, C.-Y., Bochkovskiy, A. and Liao, H.-Y.M. (2023) 'Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors', 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) doi:10.1109/cvpr52729.2023.00721.